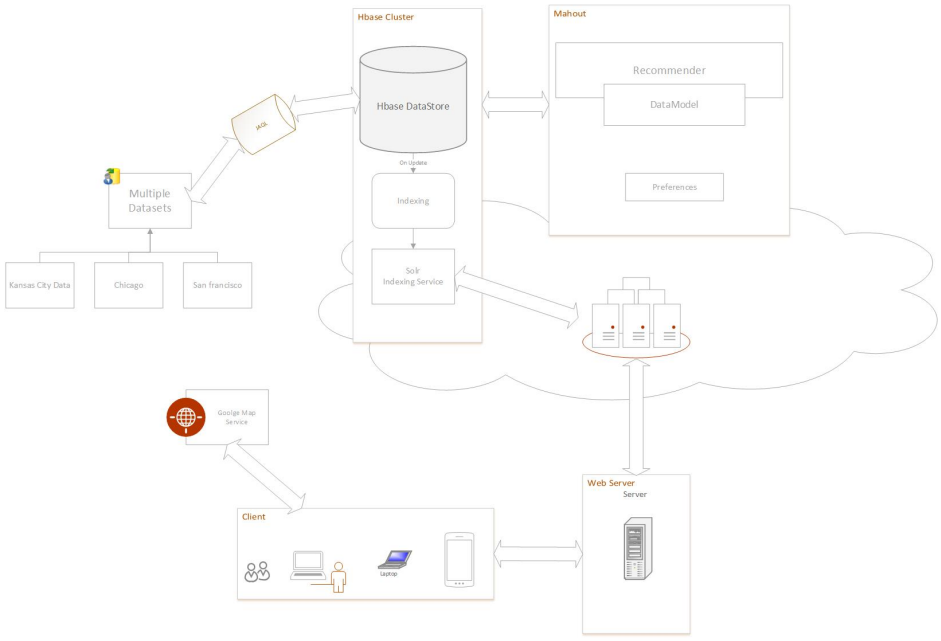
4-1-2014

SAGNAS (Group 9) Second Increment

1. **Framework Specification:**
   We've modified the framework for our application. We'll be using a Microsoft Azure server with Cloudera Edition installed. We'll utilize HBase exclusively for storage, R for data processing and machine learning and Impala for importing the R results to HBase. This removes the need for JAQL, Solr and Mahout. The client-portion remains the same.

System Architecture of SAGNAS



- **Domain Model**

  Data Sources

  We will need to store two different types of datasets:

  - The crime event data set, which can be used to display specific crime events as well as general crime statistics. This dataset is taken from city government websites, and is used to populate our analytical, R generated data.

- The R generated dataset, which will use various attributes within the crime data set to generate various analytical information that can help the user be safer.

Methodologies and Algorithms

Following steps are involved in processing the datasets using R.

1. Data Extraction
   - We use the data from following cities datasets from opendata.org community.
   a) Kansas City
   b) Chicago
   c) San Francisco

   - Data is extracted in CSV format from the source website and loaded in R

2. Data Exploration and Preprocessing
   - We need to understand how the data is organized and different fields available in the dataset.

3. Data Processing and Cleaning
   a. The data is present in terms of records with exact time, location, etc. So distributing the data w.r.t month, week, zipcode, time interval.
4. Extraction of results

Analytic Tools

Our main tool to analyze our data will be R and its data processing/machine learning algorithms library.

Analytical Tasks

R will be used for pre-processing the data in order to get it ready to be used from a machine learning standpoint. We have found utilizing R to be much more efficient than manually processing the data then inserting into Mahout.

Data Model Clustering Mechanism:

Clustering Mechanism is based on the safety of a specific region which is decided based on the number of crimes took placed over the past few years.

The safety parameter is grouped as:

- Relax - Number of Crimes < 20
- Safe - 20 < Number of Crimes < 50
- Moderate - 50 < Number of Crimes < 100

- Less Dangerous - 100 < Number of Crimes < 200
- Dangerous - Number of Crimes > 200

The decision is made based on the Number of crimes took place in the region and the time period like a specific month or year.

Prediction Parameters:

Input Parameters: Zip Code(Default Current Location), Crime Type, Duration.

Output Parameters: Based on Crime rate our system is going to help the user to predict the occurrence of Crime during a certain period.

- **Application Specification**

Features List

New Services to be built:

- User recommendation service
  - Recommendations will be calculated weekly (not real-time)
  - Calculations will take the crime event dataset as an input, and produce a user recommendation dataset as an output
  - Will notify users of rate of crime happening in a certain area
  - Will utilize an R generated recommendation dataset accessed from HBase

- Crime information service
  - The input it uses is the crime event dataset
  - Will be displayed using charts and tables, using information based on the crime event dataset stored and accessed from HBase
- Crime display service
  - Will be displayed on Google Maps, using information based on the crime event dataset stored in HBase
  - Ideally crime events would be displayed by zipcode, but due to various difficulties our plan instead is to display all events

Design of Mobile Client

We will utilize HTML5, Twitter Bootstrap, AJAX and CSS3

- **Implementation**

Implementation of data model and algorithms

Our data model stores data produced from R machine learning into HBase from Impala for easy client access. The crime data will be read from HBase after R' classification algorithm is run upon it.

We've resolved our general architecture issues, and currently working on each segment separately (R data processing/machine learning, client/HBase communication.

Implementation of services

We are going to use an external tool, R, in combination with HBase.

1. A RESTful service to import R's data to HBase via Impala
2. A Service is used to initiate the indexing process of HBase data and store the values in HBase.

Implementation of user interface

Our user interface is designed for mobile devices. We are utilizing twitter bootstrap to maintain a clean interface, and JSP to ultimately display the client-side code we have written.